Twitter Event Signatures: An investigation of how real-world events are reflected in social media signals.

Natan Grayman

School of Electrical & Information Engineering, University of the Witwatersrand, Johannesburg, South Africa

Abstract: The aim of this investigation report is to elucidate the process of exploring the intricate relationship between various event types and the corresponding patterns of interest and trends observed on Twitter. Extensive research into open-source tweet databases was undertaken, resulting in the use of datasets from platforms including Kaggle, the Internet Archive, and GDELT. The event types analysed encompassed protests, natural disasters, and election cycles, each of which involved the collection and extraction of two sets of event datasets. The analysis of each compiled dataset extended to various dimensions, including geolocation, news media response, and trend validation. Mathematical techniques such as seasonal decomposition and the LOESS regression method were employed to approximate and evaluate the trends. Ultimately, this facilitated the creation of a framework for classifying each event with a distinct event archetype. The results and key findings from this investigation offer valuable insights into the relationship between event types and Twitter trends, paving the way for more precise event classification and trend analysis.

Key words: *Twitter trends, event archetype, open-source tweet data, seasonal decomposition, LOESS regression.*

1. INTRODUCTION

In this investigation project, the primary objective is to explore the intricate relationship between various event types and the corresponding patterns of interest and trends observed on Twitter. Through this investigation the overarching question arises of how real-world event types are reflected in the nuanced signals of social media, particularly Twitter. This analysis and comprehension of social media responses to different event types can yield novel insights into online interactions and the foundations of contemporary society [1].

The following sections will guide through the key aspects of the investigation. Commencing with a background section, the concept of event signatures is explored, elucidating their manifestations and examination on Twitter, along with background of other open-source platforms. Subsequently, the investigation methodology explains the approach to restrict and define the event types under examination, as well as the methods utilized to retrieve the corresponding data.

Thereafter, the key findings and results section provides a detailed explanation of the results at each stage of the investigation project, offering insights into the discovered patterns and trends. Finally, the discussion section critically assesses the project's limitations and outlines future research prospects.

2. BACKGROUND

In this section the background on the literature on what is an event signature and the impact of the change in ownership of Twitter on this investigation project will be discussed.

2.1 Event Signatures

In the digital age, the pervasive influence of social media platforms has reshaped the way people interact with live events, enabling users to instantaneously engage, react, and participate in global discussions. Consequently, during significant societal occurrences, be they natural disasters or political upheavals, individuals often turn to these platforms as their initial response, forming a digital footprint that chronicles their reactions [1]. This digital footprint can be encapsulated by the term "Event Signatures," signifying the distinctive patterns of public interest that emerge in response to various types of events [1].

The “Event Signature’’ is a comprehensive representation of any reactive metric, depicting the evolving user engagement and interest in a particular event [2]. This comprehensive digital footprint encompasses user activities, such as posts, comments, likes, edits, reposts, and hashtag usage, collectively reflecting the degree and nature of public interest in the given event [2].

The analysis of event signatures offers invaluable insights into the inherent qualities of the events, facilitating the development of future mitigation and response strategies [1] [2].

2.2 Twitter Analysis

Twitter, the prominent microblogging platform known for its real-time information sharing, underwent substantial changes in 2023 following its acquisition by Elon Musk [3]. These changes included a rebranding to ‘’X’’ and the monetization of its API [3].

By analysing event signatures on Twitter, it becomes

possible to examine the dynamic and temporal volume of tweet reactions to a particular event, revealing the pattern of interest over time [1] [2]. This analysis is facilitated by extracting tweets based on keywords, hashtags, and dates [3]. The event signature, in this context, can be described as a numeric sequence representing the daily tweet volume:

$$S = \{s_1, s_2, \dots, s_n\} \quad (1)$$

Where:

s_i = the volume of event-related content in time period i.

2.3 GDELT

Supported by Google, the Global Database of Events, Language, and Tone (GDELT) is a dynamic and continuously updated dataset that comprehensively monitors and records a wide array of global events, news articles, and media sources worldwide [4]. GDELT is recognized as the largest open-source database of human society, making it an invaluable resource for researchers, analysts, and data scientists seeking insights into global events, trends, and sentiments spanning several decades [4].

3. METHODOLOGY

This methodology section explores data collection methods, their impact on event selection, the findings of the pattern of interest, geolocation, and trend validation.

3.1 Data Collection

Despite the constraints posed by the monetization of the Twitter API, the investigation project proceeded and overcame the issue by utilizing open-source data. Consequently, the methodology for selecting events to investigate was contingent on the availability of high-quality tweet datasets corresponding to specific event types.

In the early stages of dataset research, the Wharton and Annenberg Historical Dataset appeared promising for high-quality data [5]. Unfortunately, access to this dataset was denied. Subsequent exploration of various GitHub repositories revealed issues such as data corruption, insufficient volume, and dataset inconsistencies.

Nevertheless, the investigation resulted in the identification of a substantial data source archive known as The Internet Archive. This 501(c)(3) non-profit organization is dedicated to creating a digital library of internet sites and other cultural artifacts in digital form [6].

This dataset proved to be extensive, with an impressive volume of approximately 4,000 to 5,000 tweets recorded every minute within the specified date range [6]. However, it is important to note two specific constraints associated with this dataset. Firstly, the dataset's temporal coverage is restricted, as it does not include tweets from November 2022 to September 2023, thereby limiting the analysis of events to the earlier time frame. Secondly, the tweet data extraction method is a light tier, lacking tweet features such as longitude and latitude information.

Throughout the investigation project, data extraction from the Internet Archive was effectively refined and optimized. Initially, the archive's data required decompression and was searched using a variety of relevant keywords, hashtags, and locations. This extraction process was both tedious and error-prone, leading to errors in dataset formatting and extraction handling.

Likewise, the initial use of a local computer for extraction resulted in slow processing and high battery consumption. However, these initial challenges provided valuable learning opportunities. Subsequently, the team transitioned to establishing efficient access and usage of the Wits cluster while implementing multiprocessing for extraction. This optimized process enabled the utilization of a template code for various events, streamlining the addition of new datasets for archetype creation.

Furthermore, Kaggle, an online community platform for data scientists and machine learning enthusiasts, served as another crucial source for data collection [7]. The utilization and leveraging of high-quality data on Kaggle, combined with the flexibility of the Internet Archive, enabled the team to make well-informed decisions regarding which event types to investigate.

3.2 Event Selection

The event types selected based on the availability of ample, high-quality data were Elections, Natural Disasters, and Protests. Specifically , the events included the 2020 US election tweets from Kaggle, the 2023 Turkey and Syria Earthquake tweets from Kaggle, and the 2021 South African unrest tweets extracted from the Internet Archive.

Following the methodical extraction, cleaning, and exploration of these datasets, the investigation into signature analysis was initiated, facilitating an in-depth exploration of the distinctive pattern of interests.

3.3 Pattern of Interest

The primary goal of this investigation project, which aimed to uncover the pattern of interest for each distinct event type, was successfully achieved.

Appendix C illustrates the pattern of interests identified for each of the initial events within the Protest, Election and Natural Disaster event type with contextual date markings.

When comparing the patterns of interest, it became evident that the varying count sizes within each dataset rendered graph comparisons challenging due to their distinct scales.

To address this issue, normalization, a statistical technique employed to rescale data and bring it within a common range, was utilized [8]. It proved crucial throughout the investigation project in facilitating a fair assessment of the shapes and trends of different curves, preventing variations in magnitude or scale from overshadowing the underlying structural patterns.

3.4 Geolocation

Analysing the features of each Twitter dataset revealed key findings that deepened the understanding of the data and directed subsequent steps in the investigation. One notably interesting aspect was Twitter's ability to provide geotagging coordinates in the form of longitude and latitude features. This facilitated the creation of geographical maps for each event. However, it's important to note that this feature was exclusively available in Kaggle datasets containing geotagging information. In contrast, the Internet Archive dataset provided solely textual location descriptions gathered by Twitter. To convert these textual location descriptions into geotagging coordinates, the open-source Nominatim API was employed [9]. Figure 1 illustrates the geocoded map of reactions during the 2021 South African Social Unrest.

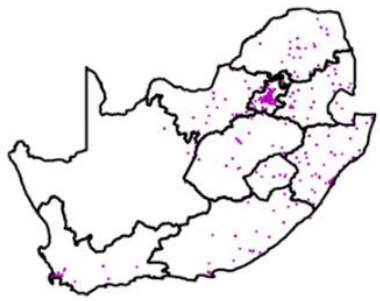

Figure 1: Geolocation map of Twitter activity during the 2021 South African Social Unrest

While addressing the challenge of missing coordinate features, this exploration led to an examination of GDELT's functionality.

3.5 News Comparison

Through GDELT's open-source data, accessible via Google BigQuery, the total number of news articles published by registered news companies during each event can be extracted. This enables a visual comparison between the shape and trend of the news coverage and social media response. Appendix D illustrates this comparison for each event type.

Using Figure 8 in Appendix D of the South African social unrest as an example, a noticeable delay can be observed, with the news media peaking in response later compared to the peak in social media activity, notably when the army was deployed.

3.6 Trend Validation

Investigating both the pattern of interest in tweets and the temporal changes in news articles over the same duration raised questions about the validity of the observed trends. This led to an exploration of additional sources, including the Mastodon API, an open source, decentralized social media platform similar to Twitter [10].

After extracting Mastodon's posts or “statuses”, it became evident that there was insufficient data available for both the South African Social Unrest and the Turkey-Syria unrest events [10]. Nevertheless, the extraction for the US election yielded ample data, allowing validation of the observed trend during the election cycle as shown in Figure 14.

Additionally, Google Trends played a crucial role in the trend validation process. Google Trends is a website that analyses the popularity of search queries in Google Search across various regions and languages, providing users with graphical trends to compare the search volume of different queries [11]. It also features clustered topics related to trending searches on Search, Google News, and YouTube, offering insights into both live and historical search trends that capture people's reactions [11]. Therefore, utilizing the trend functionality for search volume trend data for both search queries and news articles allowed the validation of the general trend of user reactions throughout each event's duration.

Using Figure 11 in Appendix E , evidently, the trend observed on Google by users searching for key terms like "Looting," "State of Emergency," and "Protest" closely follows the upward peak and gradual downward trajectory of the pattern of interest in Figure 5.

4. RESULTS AND FINDINGS

To ensure comprehensive access to all the results, the project team has generated three Kaggle Jupyter notebooks, utilized for its cloud computation capabilities and shared workspace. These notebooks cover investigations each event type: Elections, Protest and Natural Disasters.

Each Jupyter notebook adheres to a similar structure, offering a table of contents and detailed explanations. Nonetheless, this section will explain the mathematical framework and resulting event archetypes.

4.1 Mathematical Framework

To investigate the trends observed in all events, a mathematical trend analysis was conducted. Seasonal decomposition emerged as an effective framework for analysing the time-series data [8]. Seasonal decomposition entails breaking down the time-series data into three primary components [8]. The trend component reveals the underlying long-term behaviour within the data, shedding

light on whether it displays a general trend of increase, decrease, or stability over time [8]. Seasonal patterns, characterized by recurring fluctuations at fixed intervals, are captured within the seasonal component, aiding in the identification of consistent patterns in events or data [8]. The residual component encompasses unexplained or random fluctuations, proving valuable for detecting anomalies or noise not explicable by trends or seasonality [8].

Furthermore, seasonal decomposition provides two core methods, additive and multiplicative, to break down data into trend, seasonal, and residual components, with the additive approach applicable when seasonal effects remain constant over time and the multiplicative approach suitable when the seasonal influence scales with the data's level [8]. Moreover, a "period" parameter in seasonal decomposition defines the number of data points within a single season, enabling alignment with the data's inherent seasonal patterns. As a result, the exploration of seasonal decomposition involved the use of a slider widget to explore various periods and a widget to choose between additive or multiplicative methods [8].

The resulting seasonal decomposition verified the hypothesis that a small period , of a few days within each event , showed an effective aligned trend, seasonal oscillation and a relatively low residual curve. However, the results obtained from seasonal decomposition are limited by their reliance on trend curves that seek to identify long-term general patterns of growth, decline, or stability over time, rather than capturing the inherent complexity within each event type [8].

Therefore, to utilize a more granular and controlled technique for approximating the trend of each pattern of interest, the LOESS (Locally Estimated Scatterplot Smoothing) non-parametric regression technique was employed [9]. This approach offers the advantage of a smoothing parameter that can be adjusted through trial and error to effectively increase or decrease the smooth bandwidth for each pattern of interest [9]. However, despite LOESS's flexibility in modelling complex curves for which no theoretical models exist, it does not produce a regression function that is represented by a mathematical formula [9].

4.2 Event Archetype

After analysing each event, including its pattern of interest, news comparison, trend validation, and mathematical framework, the investigation progressed to explore additional data sources for inclusion. Thanks to the optimized data extraction processes, lessons learned throughout this study and streamlined code templates, the seamless integration of new datasets into each event type was facilitated.

In the Protest event type, investigations were conducted to consider the inclusion of both the Black Lives Matter Protests in 2020 and the Mahsa Amini Iran Protests. However, the Black Lives Matter tweet data examined was found to lack sufficient data for the entire event. In the Election event type, the US election cycle in 2016 was introduced. For the Natural Disaster event type, the Mexico earthquake in September 2017 was integrated. Following the incorporation of these events into the Jupyter notebook for each event type, the LOESS mathematical framework was employed to explore each trend and ultimately examine the feasibility of formulating an archetype for each event type.

The resulting archetypes, established with the use of upper and lower boundaries, represent the potential range within which the Twitter trend could be positioned for each event type. The implementation of these boundaries is essential for capturing the average trend between the curves, offering a range in which the archetype is situated and allowing for a more accurate representation of the typical behaviour of each event type.

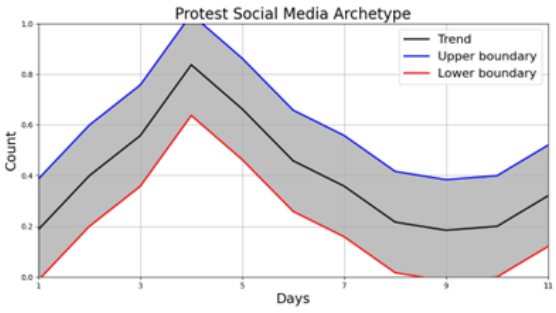

Figure 2: Event Archetype of Protest event type

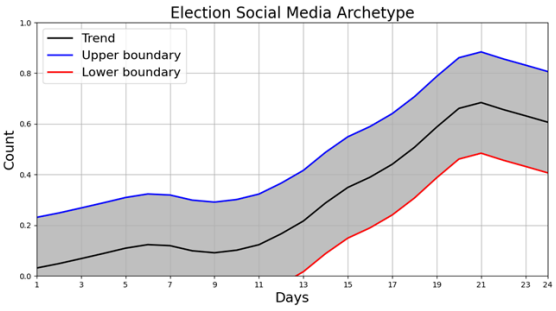

Figure 3: Event Archetype of Election Archetype

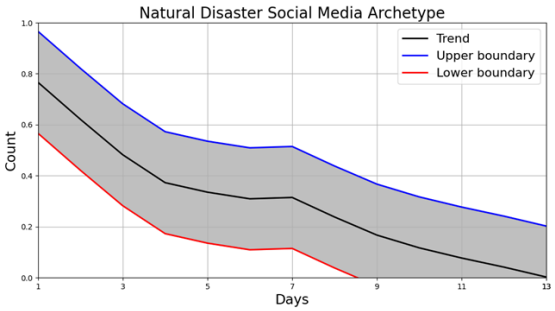

Figure 4: Event Archetype of Natural Disaster

To validate the resulting archetypes, similarity metrics were employed to assess the proximity of each pattern of interest to the overall archetype. The first metric employed is the Euclidean distance, which quantifies the spatial or geometric distance between two data points [10]. Since both curves are normalized, the Euclidean distance is suitable for measuring the magnitude of points or outliers from the archetype trend [10].

Likewise, the Mean Squared Error (MSE) metric was employed to assess the average squared differences between two datasets, offering a measure of their overall dissimilarity [11]. This metric proves suitable for comparing the archetype to the pattern of interest as it enables the evaluation of error magnitudes in trend data, facilitating a meaningful assessment of their overall fit [11]. The range of MSE values enables the quantification of how closely the pattern of interest aligns with the archetype, with lower MSE values signifying a closer match and higher values indicating greater dissimilarity [11].

Furthermore, the Cosine Similarity metric was employed to facilitate directional relationship quantification [12]. It calculates the cosine of the angle between the two compared data points and, as a result, quantifies the directional similarity between the points in each curve [12]. The scale of Cosine Similarity ranges from -1, signifying complete dissimilarity, to 1, indicating perfect similarity, while a value of 0 suggests that the data points are unrelated or uncorrelated [12]. Similarly, to explore the linear directional relationship between the archetype and each pattern of interest, the Pearson correlation coefficient was employed [13]. To illustrate these results, Table 1 shows the resulting similarity metrics for the Election event type.

Another mathematical feature investigated was the use of cross-correlation to quantitatively model the time delay between social media and news media responses. By treating the response curves as signals and employing convolution, cross-correlation identifies the time lag at which the two signals exhibit maximum similarity or divergence [14]. The cross-correlation formula for two discrete datasets X and Y, each with n data points, is expressed as:

$$R(k) = \sum(X(t) \cdot Y(t-k)) \quad (2)$$

Where:
R(k) = denotes the cross-correlation at time lag k [14].

Table 1 showing the similarity metrics comparing the US election pattern of interests to the event archetype.

Pattern of Interest	Euclidean Distance	Cosine Similarity	Pearson Correlation	Mean Squared Error
US Election 2020	1.2998	0.8759	0.8177	0.0298
US Election 2016	0.5350	0.9762	0.9487	0.0309

5. DISCUSSION

Table 1 illustrates the similarity between the patterns of interest associated with the US election cycles and the event archetype. This is evident from the low Euclidean Distance, low MSE, high Cosine Similarity, and high Pearson Correlation, which collectively validate this close resemblance. These findings have significant implications, as they indicate that the event archetypes could successfully classifies reactions on social media.

5.1 Limitations

The investigation project encountered an initial obstacle when changes in Twitter's ownership structure rendered Twitter's API inaccessible as a data source. This inaccessibility was primarily due to the significant financial resources required to handle the extensive data volume involved in this project, which exceeded a terabyte.

Consequently, the utilization of open-source Twitter datasets in this research introduced several notable limitations. Firstly, these datasets could exhibit inherent biases that could predominantly reflect specific demographics or characteristics of Twitter users, potentially constraining the generalizability of findings. Moreover, the presence of potential data fabrication or bot-generated content within open-source Twitter datasets could raise concerns about the authenticity and reliability of the information.

Finally, these limitations also extended to the scope of the data, as open-source datasets, such as the Internet Archive, often offered only a partial representation of the broader Twitter landscape, risking the omission of tweets or trends that could impact research outcomes.

5.2 Future Research

The outcomes of this investigation project have laid a robust foundation for potential future research endeavours. The establishment of a comprehensive template for extracting tweets from reputable Twitter archives, exemplified by the Internet Archive, along with the subsequent procedures for data analysis ,cleansing and mathematical formulations, has not only streamlined the investigative process but also exemplified a heuristic approach that can be replicated and extended to encompass additional event types.

As the methodology becomes further refined and event data accumulates, this template may serve as a valuable tool, analogous to a 'sensor,' enabling real-time Twitter monitoring and event forecasting, specifically in the context of protests, natural disasters, or election cycles. By leveraging the event archetype, future research can delve into forecasting mechanisms and user interaction analysis with a heightened level of precision and accuracy, offering novel insights into these critical aspects of contemporary society.

6 CONCLUSION

In conclusion, this investigation aimed to shed light on the process of exploring the intricate relationship between diverse event types and the corresponding patterns of interest and trends observed on Twitter. Extensive research was conducted, leveraging open-source tweet databases from platforms such as Kaggle, the Internet Archive, and GDELT. The study focused on three key event types: protests, natural disasters, and election cycles, each involving the collection and extraction of two sets of event datasets. The comprehensive analysis of each compiled dataset extended across multiple dimensions, including geolocation, news media response, and trend validation. To approximate and assess the trends within these datasets, mathematical techniques like seasonal decomposition and the LOESS regression method were thoughtfully employed. The culmination of these efforts resulted in the creation of a framework for classifying each event with a distinct event archetype. The invaluable insights gleaned from the results and key findings in this investigation offer a deeper understanding of the intricate relationship between event types and Twitter trends, setting the stage for more precise event classification and trend analysis in the future.

REFERENCES

[1] H. M. Saleem, Y. Xu and D. Ruths, “Effects of disaster characteristics on Twitter event signature.,” ScienceDirect, Montreal, 2014.

[2] J. L. Jaewon Yang, “Patterns of temporal variation in online media,” WSDM '11: Proceedings of the fourth ACM international conference on Web search and data mining. ACM Digital Library, 2011.

[3] X Developer Platform, “X API,” 2023. [Online]. Available: https://developer.twitter.com/en/products/twitter-api . [Accessed 23 October 2023].

[4] GDELT, “The GDELT Project,” [Online]. Available: https://www.gdeltproject.org/about.html. [Accessed 24 October 2023].

[5] Wharton University of Pennsylvania, “Dataset of Historical Tweets,” [Online]. Available: https://research-it.wharton.upenn.edu/data/tweet-database/. [Accessed 23 October 2023].

[6] Internet Archive, “About the Internet Archive,” [Online]. Available: https://archive.org/about/. [Accessed 23 October 2023].

[7] Kaggle, “Datasets,” [Online]. Available: https://www.kaggle.com/datasets. [Accessed 23 October 2023].

[8] R. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice (2nd ed), Monash University, Australia: Otexts, 2023, pp. 156-238.

[9] W. S. Cleveland, “LOWESS: A Program for Smoothing Scatterplots by Robust Locally Weighted Regression,” The American Statistician, Vol. 35, No. 1 (Feb., 1981), p. 54 (1 page), 1981.

[10] NIST, “Euclidean distance,” [Online]. Available: https://xlinux.nist.gov/dads/HTML/euclidndstnc.html . [Accessed 25 October 2023].

[11] Britannica, “mean squared error,” [Online]. Available: https://britannica.com/science/mean-squared-error. [Accessed 25 October 2023].

[12] F. Karabiber, “Cosine Similarity,” LearnDataSci, [Online]. Available: https://www.learndatasci.com/glossary/cosine-similarity/. [Accessed 25 October 2023].

[13] KentState University, “SPSS tutorials: Pearson correlation,” [Online]. Available: https://libguides.library.kent.edu/spss/pearsoncorr. [Accessed 25 October 2023].

[14] All About Circuits, “Understanding Correlation,” [Online]. Available: https://www.allaboutcircuits.com/technical-articles/understanding-correlation/. [Accessed 25 October 2023].

[15] Nominatim Manual, “Overview,” Nominatim developer community, 2023. [Online]. Available: https://nominatim.org/release-docs/develop/api/Overview/. [Accessed 24 October 2023].

[16] Mastodon, “What is Mastodon?,” [Online]. Available: https://docs.joinmastodon.org/. [Accessed 24 October 2023].

[17] Google, “Google Trends,” [Online]. Available: https://trends.google.com/trends/. [Accessed 24 October 2023].

Appendix A: Group Work Reflection

Introduction

The group work undertaken by Liad Peretz and I set out to investigate how real-world events are reflected on social media. This project was driven by our shared interests in Information Engineering and our strong compatibility as partners. Our mutual curiosity greatly facilitated our project decisions, allowing us to utilize each other's strengths to effectively problem-solve throughout each project stage. Our research aimed to uncover the fascinating ways in which real-world events are mirrored in the digital realm, specifically on Twitter. This exploration not only holds academic significance but also practical relevance in the realm of Information Engineering, where understanding the dynamics of social media platforms is crucial. Our partnership was founded on the fusion of these objectives and our individual passions, setting the stage for a collaborative journey in which we each brought distinct skills and perspectives to the table.

Overview of Group Activities Throughout the Project

Throughout the project, our group's collaborative efforts led to a well-balanced distribution of tasks, fostering effective communication and coordinated time management. We approached our work with a strategic division of responsibilities for each project stage and maintained a consistent schedule of meetings, either in-person or online, tailored to the project's evolving needs. Furthermore, we conducted daily or weekly reviews of each other's work to ensure quality and progress. In addition, our weekly meetings with our supervisor played a pivotal role in obtaining key insights, valuable advice, and efficient planning for the upcoming week, which we successfully implemented. This collaborative framework not only enhanced our productivity but also contributed significantly to the project's overall success.

In the project's initial stage, a significant setback arose due to changes in Twitter's ownership structure, rendering the Twitter API inaccessible as a data source. In response to this unforeseen obstacle, I assumed the lead role in researching alternative open-source data sources and meticulously evaluating their quality, particularly with respect to the diverse event types we were investigating. This extensive research culminated in our decision to utilize Kaggle, the Internet Archive, and GDELT as our primary data sources, judiciously chosen to align with the unique requirements of each stage of our investigation. This adaptive approach allowed us to navigate and overcome the unexpected challenge, ensuring the project's continued progress and success.

Upon completing our evaluation and satisfaction with the selected data sources, we honed our focus to three specific event types: Protests, Natural Disasters, and Elections, marking the commencement of our data analysis phase. A critical aspect of this stage was the meticulous extraction and cleaning of the data.

To facilitate the setup and configuration of data extraction from the Internet Archive, GDELT, and later the Wits Cluster, Liad and I collaborated closely. We jointly deliberated on the concept, logic, environment configuration, and code structure, ensuring a well-coordinated approach. After establishing the initial infrastructure, we strategically divided the substantial workload and assigned tasks based on our strengths.

Liad took the lead in optimizing and refining the data extraction processes, ensuring the data's integrity and quality. Meanwhile, I immersed myself in the exploration of the available datasets, leveraging the collaborative features of Kaggle's Jupyter notebook environment, which we chose for its non-local, shared workspace. Within this environment, I assumed responsibility for configuring and structuring the notebook, an essential step in contextualizing each event type. This contextualization served as the foundation for unravelling the narratives concealed within the data. As part of this process, I delved deep into the data, seeking to uncover insights and patterns. Specifically investigating features such as daily or hourly tweet count ,hashtags, retweets, like counts, location metrics etc. This meticulous examination enabled us to gain a comprehensive understanding of the events under scrutiny, contributing significantly to the quality of our analysis and the depth of our findings.

Building upon our accomplishments in the extraction and analysis stage, our collaborative efforts led to the successful achievement of our primary goal: the creation of contextual patterns of interest for each event type. With this milestone behind us, we seamlessly transitioned into the next phase of our investigation, which involved raising questions about the data findings.

Firstly, we critically examined the patterns of interest we had meticulously identified and delved into understanding how these findings resonated with responses in the news media. This comparison between social media trends and news media responses was a crucial aspect of our research. Leveraging my research on GDELT and Liad's optimization of data extraction from GDELT's BigQuery platform , we successfully conducted these comparisons for each event.

Furthermore, as we delved into the analysis of the event signatures, we became curious about the nature of the trends we had uncovered. We questioned whether they were valid, indicating true representations, or anomalies. To address this, Liad focused on data extraction from the Mastodon API, a Twitter competitor. However, due to the limited user base on Mastodon, we found sufficient data only for the US Election. Concurrently, I explored alternative methods to validate these trends and opted for Google Trends, which offers both web search and news search functions. Utilizing Google Trends, I extracted search and news queries for each event, enabling a direct comparison across all events. This approach provided validation for the trends we observed in both social media and GDELT articles.

Moreover, we delved into the exploration of trends and mathematical approximations derived from the data. Guided by Dr. Martin Bekker, we explored Seasonal Decomposition and LOESS regression techniques, collaborating closely to mathematically characterize each event. During this phase, I focused on the investigation of similarity techniques, including autocorrelation and cross-correlation, while also optimizing the outcomes from the seasonal decomposition process. Simultaneously, Liad expanded data retrieval efforts for each event type, a crucial step toward creating the archetype for each event.

Ultimately, through our collaborative efforts, starting from the initial research stage and culminating in effective data synthesis, enabled us to craft a distinctive archetype for each event type.

Conclusion

In conclusion, the investigation project undertaken by Liad Peretz and I resulted in a successful investigative project. Despite initial setbacks, our robust group collaboration propelled us beyond the initial goal of analysing patterns of interest. We delved deeper into the data, ultimately establishing event archetypes for not one but three distinct events, involving an extensive volume of data exceeding one terabyte. The success of this endeavour was made possible through the effective partnership between Liad Peretz and I, as well as the invaluable guidance and support provided by our supervisor, Dr. Martin Bekker.

Appendix B: Project Plan

Title: Investigation Project Proposal Plan- Digital Event Signatures

Liad Peretz (2373287) and Natan Grayman (2344104)

School of Electrical & Information Engineering, University of the Witwatersrand, Private Bag 3, 2050, Johannesburg, South Africa

Abstract: This project proposal plan aims to represent the strategy to investigate patterns of interest in digital event signatures and their relationship to different types of events. A succinct literature review and investigation into existing solutions provided valuable insights into digital signature extraction methods and influencing factors. The methodology proposed involves data collection, data preprocessing, event signature creation, modelling and analysis and the compilation of key findings for the documentation and presentation. A proposed project timeline, activity and milestone schedule has been created to ensure an effective structure and planned progression throughout the project. The inclusion of the work breakdown structure, methodological flow chart, and risk register enhances the proposed planning by providing a structured breakdown of tasks, illustration of the logical flow of activities, and outline of contingencies to address potential risks. Collaboration, balanced workload allocation between team members and ethical considerations are planned to be upheld throughout the project. Ultimately, the findings from this research have the potential to contribute to event categorization and understanding based on digital signatures.

Key Words: Event Signature, API, web scrapping, data collection, modelling

1. INTRODUCTION

1.1. Project Specifications

This project aims to investigate the relationship between different types of events and the corresponding patterns of interest observed in their digital signature.

Within the scope of this project, an event is to be considered as a specific incident or subject of interest that generates user responses and activities on social media or other digital platforms. The event could incorporate a wide range of topics, such as social, political, cultural, or environmental occurrences that capture public attention and prompt individuals to engage and express their thoughts, emotions, or support in a digital format.

During significant events, such as major societal developments, natural calamities, political shifts, or noteworthy incidents, individuals often resort to social media platforms as a means to access information and actively participate in discussions related to the event. As a result, a distinct digital footprint emerges on social media platforms during these events. Consequently, “event signatures” can be identified, representing the unique patterns of public interest for different types of events [1].

Accordingly, within the scope of this project, a digital event signature refers to the distinctive pattern of user-generated content, engagements, or interactions generated on public accessible social media platform(s). It is characterized by utilizing specific metrics to represent the interest over time for the selected event [2]. Therefore, it encompasses the collective digital footprints left by users, for instance posts, comments, likes, edits, retweets, hashtags, and other relevant indicators, which reflect the level and nature of public interest in a given event.

Ultimately, analysing the characteristics of event signatures can provide valuable insights into the nature of the events themselves [3]. By investigating the unique patterns of interest associated with different events, this project aims to uncover insights into the underlying characteristics, classifications and resulting mathematical relationships of events in the digital realm.

1.2. Assumptions, Constraints and Success Criteria

In order to effectively plan and execute this project, it is important to consider the assumptions, constraints and success criteria that may impact its progress and outcomes.

Firstly, in this project, it is assumed that the user identity captured on social media platforms or through web scraping techniques is unique, allowing for accurate analysis and interpretation of the digital footprints. In cases where platform-specific metrics may affect the uniqueness of user identities, appropriate measures will be taken to validate the data.

Furthermore, it is assumed that the necessary public data required for analysis throughout the project's duration, or the specified data collection period will be cost-free. Likewise, it is assumed that the data to be collected from the internet has already undergone ethical considerations by users when they voluntarily uploaded their responses and data to social media platforms.

The project is constrained by the availability and limitations of free platform APIs (Application Programming Interfaces) provided by social media platforms like Twitter or other relevant sources. These APIs allow access to specific data and functionalities, and the project's data collection efforts are subject to any

restrictions or changes imposed by the platform providers [4].

Moreover, the quality and availability of the project's data are contingent upon the accessibility and reliability of existing open-access data sources. The project team's ability to collect comprehensive and accurate data is reliant on the availability and quality of the data already accessible from public sources. Any limitations, gaps, or inconsistencies in the existing data may impact the overall quality and validity of the project's analysis and findings.

Additionally, the project success criteria are as follows:

1. Obtain an adequate quantity of relevant data to create meaningful signatures.

2. Develop a program to facilitate data extraction.

3. Model event signatures using the collected data.

4. Use an agile approach to acquire signatures for a minimum of two event categories.

5. Compare and analyse the differences and similarities between the selected event categories.

6. Conduct an in-depth analysis to understand the factors influencing the shape of the event signatures.

2. BACKGROUND

2.1. Literature Review

Event signatures have received limited attention in the existing literature, but there are relevant research findings that can contribute to this study. Notably, the work of Haji Mohammad Saleem et al. (2014) sheds light on how factors such as event duration, severity, foreknowledge, and news media engagement influence event signatures [1]. Additionally, their study highlights the valuable insights that can be gained from analysing the data sources during the peak of event signatures [1]. For instance, it may be observed that the peak of a signature primarily consists of interests from news companies, indicating a news-driven signature peak [1]. The research article utilized the Twitter social media platform, leveraging event-related hashtags to construct the event signature [1]. The findings show that several factors, including the timing, preexisting knowledge, duration, severity, and media coverage of an event, can influence an event signature [1].

2.2. Existing Solutions

In the pursuit of tracking terrorism, Syed Toufeeq Ahmed et al. (2009) explored the utilization of event signatures. Their work provided valuable insights into extracting event signatures from news articles, offering pertinent findings that can contribute to this research. These insights encompass the extraction of word types to identify the nature of events discussed in an article, where the presence of violent words like "bombing" or "kidnapping" indicates a focus on terrorist attacks [2]. Another noteworthy revelation is that utilizing the title and the initial two paragraphs of a news article proved more effective for classification purposes compared to employing the entire article [2]. Although not specifically geared towards Twitter, these discoveries bear significance in the potential trajectory of this research, particularly in scenarios where the Twitter API may be inaccessible.

3. METHODOLOGY

The methodology encompasses data collection, data preprocessing, event signature creation, and subsequent analysis stages, employing an agile approach to effectively model and comprehend the distinct characteristics of the event signatures. The methodology is visually depicted in the Work Breakdown Structure (WBS), as illustrated in Figure 1, offering a graphical representation that provides a comprehensive and visual overview of the process. Likewise, the methodology's sequential order and components are visually represented in the flow chart provided in Figure 2 in the appendix.

3.1. Data Collection

Originally, the research proposal was focused on using the Twitter API but due to the volatility of the accessibility to the Twitter API, a conscious decision was made to adjust the project scope to investigating any online platform or dataset repository that generates and contains digital event signatures.

The data collection process begins with the development of a comprehensive plan. The data collection plan begins by identifying data sources that are accessible and available based on open access and price considerations. This includes researching and selecting sources such as social media platforms, publicly available datasets, or other relevant sources that provide access to digital footprints and user-generated content related to the events of interest.

Thereafter, research will be conducted to determine the most suitable data collection methods. This may involve utilizing API integrations, which are sets of rules and protocols that allow different software applications to communicate with each other, facilitating data extraction from platforms or services that offer APIs [5]. Additionally, web scraping techniques can be employed, which involve automated extraction of data from websites using specialized tools or data extraction scripts [6].

Once the data collection methods are determined, the next step in the plan is to identify key data variables and metrics. This comprises defining the specific data elements that need to be collected to capture the desired information

and the required frequency correspondingly required to capture real-time or historical data.

Subsequently, the data collection process will be implemented, with the chosen data collection methods set up with the required tools, scripts, or systems to capture the identified data elements from the selected sources. Furthermore, during the data collection stage, consistent monitoring will be conducted to ensure that the process stays on track and adheres to the defined timeline requirements.

3.2. Data Pre-Processing

The data pre-processing stage encompasses several key steps to ensure the quality and relevance of the data used for analysis. Firstly, the removal of irrelevant data and noise is essential. This involves employing automated or manual processes to filter out data that does not contribute meaningfully to the project, such as duplicate entries, outliers, or inconsistent data points that could distort analysis results. Additionally, techniques like word type searches, as described in section 2.2, could be applied to classify data points, and eliminate irrelevant information.

The subsequent step involves data classification and formatting, where automated approaches are employed. Data classification algorithms, machine learning models, or rule-based systems are utilized to categorize the data based on the specific criteria. Moreover, the capabilities of Large Language Models (LLMs) are leveraged to extract pertinent information and transform it into the desired format for creating signatures and conducting signature analysis [7].

Following data classification and formatting, data transformation for signature analysis is performed. This step involves developing processes and alterations to prepare the pre-processed data for accurate and insightful signature analysis. Techniques like feature engineering, normalization, or scaling are applied to transform the data into a suitable format for subsequent analysis [8]. The focus here is to maintain data integrity, quality, and relevance while mitigating the impact of noise.

3.3. Event Signature Creation

The event signature creation process involves several key steps to formulate, create, and visualize event signatures. Firstly, an initial hypothesis or assumption is formulated, taking into account domain research and preliminary insights. This hypothesis defines the expected patterns or characteristics of the event signature. Secondly, the defined hypothesis is applied to create an initial version of the event signature using the data collection and data pre-processing mentioned above. This initial signature serves as a starting point for further refinement and validation.

Moving forward, the event signatures are visualized to facilitate comparison and analysis. To identify patterns and measure the level of interest over specific durations, time bins could be employed to quantify the interest within each interval. The comparison of different time intervals helps determine the most effective bin size for the data.

Furthermore, visual representations, for instance histograms or line graphs, are created to portray the event signatures. These visualizations will provide a straightforward and accessible means of analysing and modelling the patterns and characteristics within the event signatures. These visualizations offer a straightforward and accessible means to compare and analyse the signatures.

3.4. Analysis and Modelling

The analysis and modelling stage involves key steps for interpretation, modelling, and refinement.

Initially, the event signatures undergo thorough analysis to gain a deep understanding of the patterns, features, comprehension of characteristics, and dynamics within the data. This analysis is instrumental in revealing patterns and features within the event signatures. This includes building mathematical models and conducting statistical analysis to portray the data and extract any recurring patterns, trends, or anomalies within the signatures. The mathematical analysis aims to ascertain whether events can be accurately categorized and compared to the event signatures.

Subsequently, the proposed methodological approach prioritizes feedback and iteration, fostering an agile framework that greatly benefits the project's progression and outcomes. The initial event signature is compared with expected patterns and insights, and feedback from the data is collected to validate or refine the initial hypothesis. Adjustments are made to the signature creation process based on the feedback, resulting in improved mathematical models and insights. This iterative approach ensures continuous enhancement of the event signatures and their accuracy.

The analysis and modelling stage may further include a refinement and optimization phase, where areas for improvement are identified and addressed. The models derived from each event category are compared to examine the characteristic differences among the signatures. Advanced modelling techniques, additional data sources, or sophisticated statistical methods can be explored to refine the analysis and modelling process. This iterative and optimized approach contributes to a more precise and insightful interpretation of the event signatures.

4. PROJECT MANAGEMENT

4.1. Work Schedule and Workload Allocation

Both group members will collaborate closely throughout the project, sharing responsibilities and ensuring timely completion of tasks. The work schedule will be divided per

weekly basis, focusing on the outlined methodology of data collection, preprocessing, event signature creation, analysis, and modelling. Both team members will contribute equally to all project phases, conducting research, implementing methodologies, and analysing results. Regular meetings will be held to review progress, exchange ideas, and address any challenges. By maintaining effective communication and a balanced workload allocation, the project will progress smoothly and meet the established milestones.

4.2. Timeline and Key Milestones

The tables below provide an overview of the initial timeline for the first and second block of the semester in the project, outlining the activities and milestones to be completed each week. It is important to note that the timeline is subject to change and will be updated and adjusted based on the project's velocity and trajectory. Additionally, both schedules were created with reference to the WBS (*Figure 1*) and the methodology's flow diagram (*Figure 2*).

The activities proposed for the first block include defining hypotheses, creating data collection plans, researching API documentation, exploring repository searching techniques, and preparing for the upcoming data collection stage. Flexibility in the timeline allows for adaptation and optimization as the project progresses.

Table 1: The proposed block one schedule

Week and Date	Activity	Milestone
Week 1: July 17th – July 21st	-Define minimum of 2 initial event signature hypotheses	Submit Project Proposal
Week 2: July 24th - July 28th	-Identify relevant data sources for each hypothesis	
Week 3: July 31st - August 4th	-Conduct research on the required API documentation or techniques for data retrieval	
Week 4: August 7th - August 11th	-Familiarize with web scraping techniques for data collection	
Week 5: August 14th - August 18th	-Determine key data variables and metrics for each hypothesis	
Week 6 August 21st – August 25th	-Define required data collection frequency for each hypothesis	Defining event hypotheses and data collection plans

Likewise, the table below provides an overview of the schedule for the second block project timeline. It includes the week breakdowns, activities and corresponding milestones for data collection, data preprocessing, event signature creation, analysis and modelling and compilation of documentation, allowing for a structured and comprehensive progression of the project.

Table 2: The proposed block two schedule

Week and Date	Activity	Milestone
Week 1: September 4th - September 8th	-Implement and monitor data collection schemes	-Beginning of Block 2. -Data Collection Completion
Week 2: September 11th - September 15th	-Initiate data preprocessing steps, including noise removal, data classification transformations etc	
Week 3: September 18th - September 22nd	-Finish Data Preprocessing -Visualize event signature data, preparing categorization techniques	Finalize Data Pre-Processing
Week 4: September 25th – September 29th	-Create categorization and comparison of event signatures, implementing iterative feedback process	Finish Final Signature Creation
Week 5: October 2nd – October 6th	-Build mathematical models -Continue categorization and comparison with iterative approach	
Week 6: October 9th - October 13th	-Refine and optimize results utilizing an agile approach	Finish Modelling and Analysis
Week 7: October 16th – October 20th	-Optimize, refine, and finalize results	Open Day (October 19th)

Week 8: October 23rd - October 27th	-Compilation of key findings for documentation and visualizations	Submit individual reports (27th of October)

4.3. Risk prevention and mitigation

Effective risk identification, prevention, and mitigation play a vital role in ensuring project success. A comprehensive risk register, such as the one provided in the appendix (Table 3), serves as a valuable tool for systematically documenting and tracking identified risks, their potential consequences, and corresponding mitigation strategies. Regularly reviewing and updating the risk register throughout the project lifecycle enables continuous risk management and fosters a proactive approach to addressing uncertainties and challenges.

5. ETHICAL CONSIDERATIONS

This project is committed to adhering to ethical protocols in relation to the collection and usage of social media data. It intends to meet the ethical waiver requirements specified by the Human Research Ethics Committee (Non-Medical) and School Ethics Committees, ensuring compliance with relevant laws, regulations, and ethical research codes. The methodology, tools, and resources utilized will be transparently communicated, with a focus on avoiding plagiarism and maintaining ethical standards. Moreover, the data to be collected from the internet has already undergone ethical considerations by users who voluntarily uploaded their responses and data to social media platforms, aligning with the platforms' terms of service.

6. CONCLUSION

This project proposal plan aims to represent the strategy to investigate patterns of interest in digital event signatures and their relationship to different types of events. A succinct literature review and investigation into existing solutions provided valuable insights into extraction methods of digital signatures and influencing factors. The methodology proposed involves data collection, data preprocessing, event signature creation, modelling and analysis and the compilation of key findings for the documentation and presentation. A proposed project timeline, activity and milestone schedule has been created to ensure an effective structure and planned progression throughout the project. The inclusion of the work breakdown structure, methodological flow chart, and risk register enhances the proposed planning by providing a structured breakdown of tasks, illustration of the logical flow of activities, and outline of contingencies to address potential risks. Collaboration, balanced workload allocation between team members and ethical considerations are planned to be upheld throughout the project. Ultimately, the findings from this research have the potential to contribute to event categorization and understanding based on digital signatures.

REFERENCES

[1] H. M. Saleem, Y. Xu and D. Ruths, "Effects of disaster characteristics on twitter event signature," *Procedia engineering*, vol. 78, pp. 165--172, 2014.

[2] S. T. Ahmed, R. Bhindwale and H. Davulcu, "Tracking terrorism news threads by extracting event signatures," in *2009 IEEE International Conference on Intelligence and Security Informatics*, 2009.

[3] Y. Yang, T. Pierce and J. Carbonell, "A study of retrospective and on-line event detection," 1998.

[4] B. I. Davidson, D. Wischerath, D. Racek, D. A. Parry, E. Godwin, J. Hinds, D. van der Linden, J. F. Roscoe and L. Ayravainen, *Social Media APIs: A Quiet Threat to the Advancement of Science*, PsyArXiv, 2023.

[5] D. P. Giakatos, P. Sermpezis and A. Vakali, *PyPoll: A python library automating mining of networks, discussions and polarization on Twitter*, 2023.

[6] G. Barbera, L. Araujo and S. Fernandes, "The Value of Web Data Scraping: An Application to TripAdvisor," *Big Data and Cognitive Computing*, vol. 7, p. 121, 2023.

[7] X. Deng, V. Bashlovkina, F. Han, S. Baumgartner and M. Bendersky, "LLMs to the Moon? Reddit Market Sentiment Analysis with Large Language Models," in *Association for Computing Machinery*, New York, NY, USA, 2023.

[8] J. M. Bland and D. G. Altman, "Statistics notes: Transforming data," *Bmj*, vol. 312, p. 770, 1996.

Appendix:

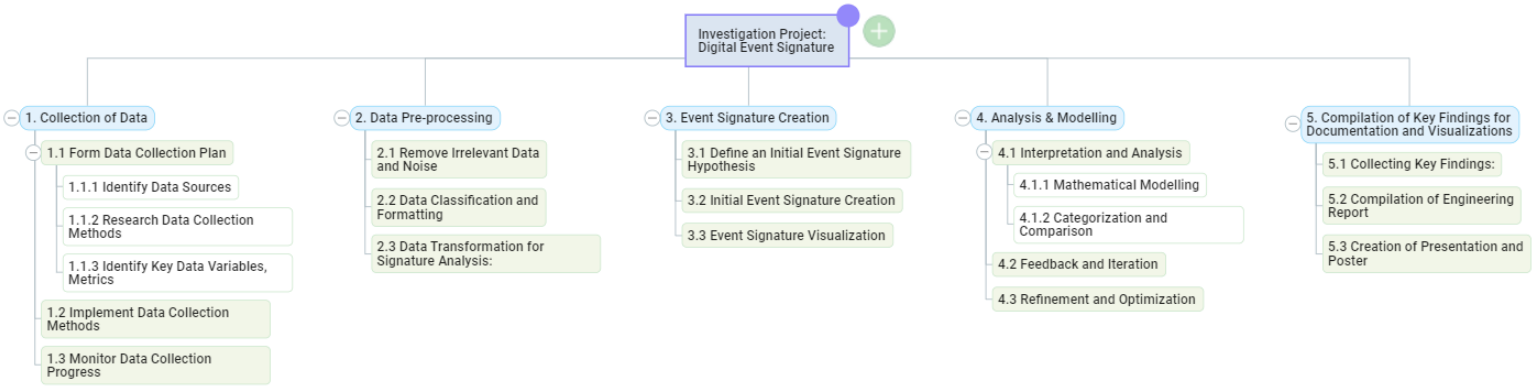

Figure 1: Work Breakdown Structure for the Project

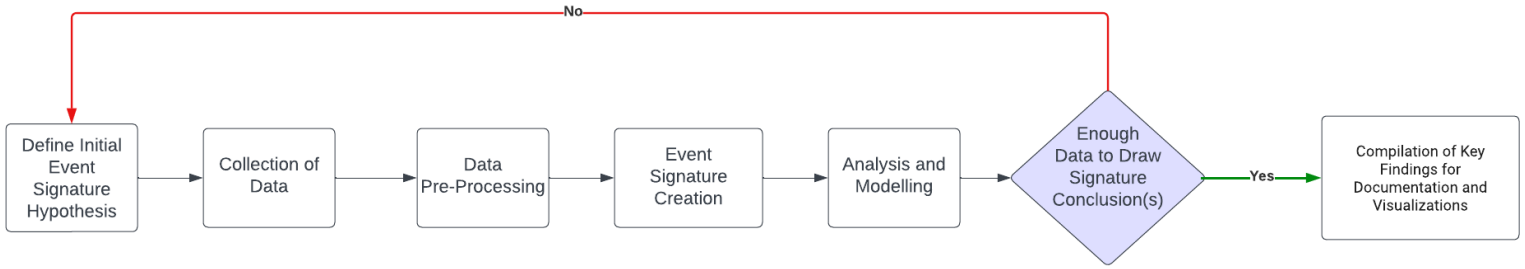

Figure 2: Flow Diagram representing Methodological Approach

Table 3: Risk Register: Identifying and Mitigating Project Risks

Risk Register						
Risk	**Causes (due to …)**	**Probability**	**Impact**	**Risk Rating**	**Response**	**Actions**
Platform policy changes	The accessibility criteria of the Twitter API or another data source have been modified, resulting in the unavailability of data from that particular source.	Medium	High	Medium	Mitigate	To enhance the chances of detecting changes in data source accessibility, it is advisable to stay updated with news and updates regarding these sources. Additionally, it is recommended to have a minimum of two alternative data sources available as backups in case the primary data source becomes inaccessible.
Data being leaked	Insufficient data protection.	Low	Low	Low	Prevent	Store data on computers that are password protected, ensuring that only authorized researchers have access to the passwords for these computers.
Low data quality	The data sources utilized exhibit inconsistencies and provide limited relevant data points.	Medium	High	High	Mitigate	When a single platform offers limited high-quality data, combining two or more platforms can be beneficial in creating a more extensive and reliable database. Alternatively, exploring additional data sources can help in finding a sufficient amount of high-quality data. It is also important to take measures to exclude bot or fake accounts from participating in the dataset.
Ethical breaches	Insufficient adherence to ethical practices, lack of oversight, and inadequate training.	Low	Low	Low	Prevent	The team participated in an ethics lecture to familiarize themselves with the protocols and guidelines for gathering data from social media platforms.
Data loss	The computer or storage containing all the necessary research data has been lost, misplaced or stolen.	Medium	High	Medium	Prevent	It is advisable to regularly create backups of data and store them on an external hard drive, separate from other storage devices that contain the same data. Additionally, storing data in the cloud provides an additional layer of protection, ensuring access to the data even in the event of issues with local storage.
Unfinished research project	The project unexpectedly involves high task complexity, necessitating the development of sophisticated tools to ensure its successful execution.	Low	Low	Low	Prevent	In order to achieve efficient progress in the project, it is essential to conduct thorough research and identify relevant techniques. Drawing upon past research or techniques can offer valuable insights and guidance for successfully executing the complex tasks required. Being proactive and making daily advancements are vital to ensure a steady and efficient project progression.
Scope Creep	The aim and requirements of the project are inadequately defined.	Medium	High	Medium	Prevent	It is important to establish clear definitions for the success criteria, goals, and the required methodology to achieve those goals.

Appendix C: Patterns of Interest

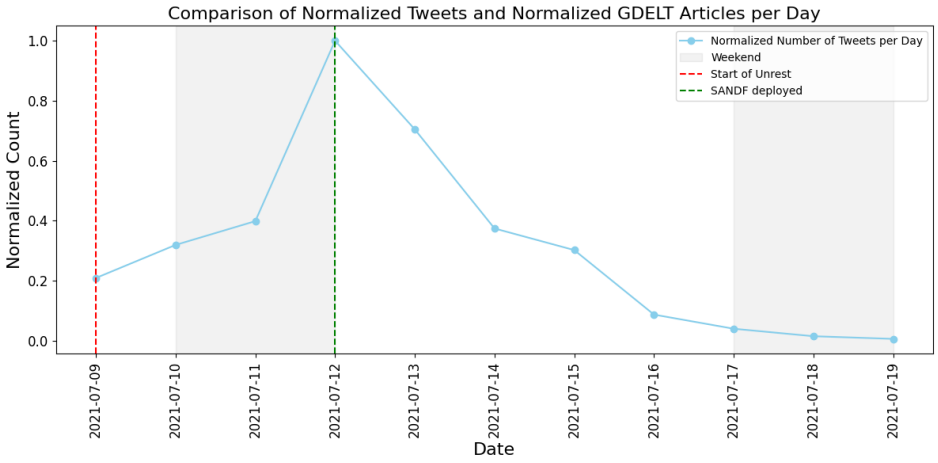

Figure 5: 2021 South African Social Unrest pattern of interest

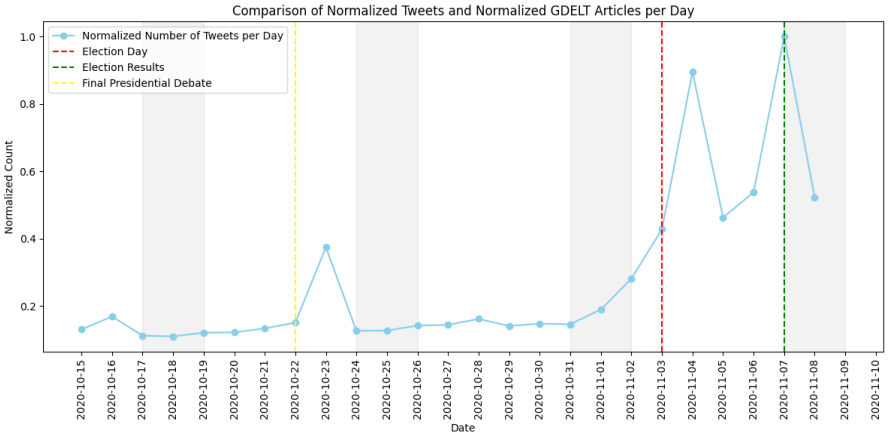

Figure 6: US Election 2020 Patter of Interest

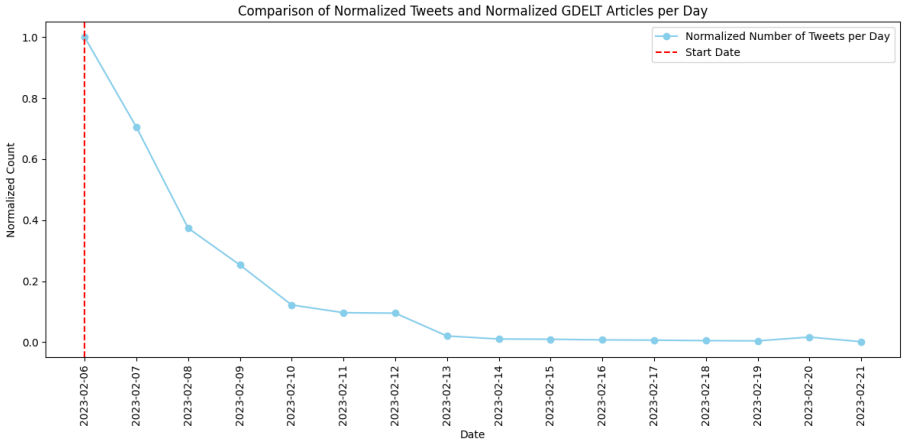

Figure 7: Turkey-Syria Pattern of Interest

Appendix D: Social Media Response compared News Media Response

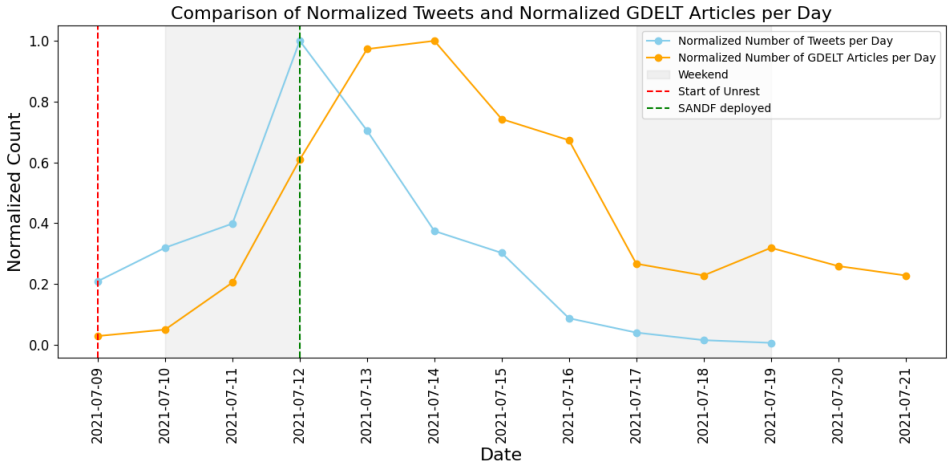

Figure 8: Comparison of social media and News Response of 2021 South African Social Unrest.

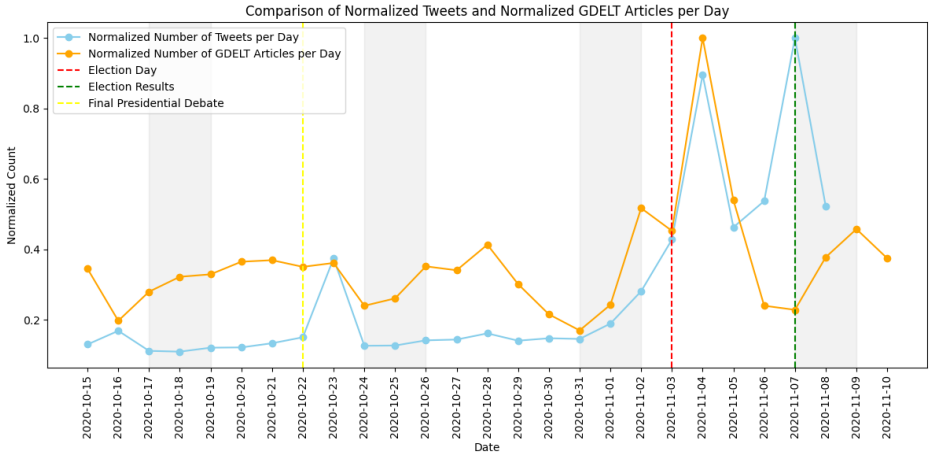

Figure 9:Comparison of social media and News Response of US Election 2020.

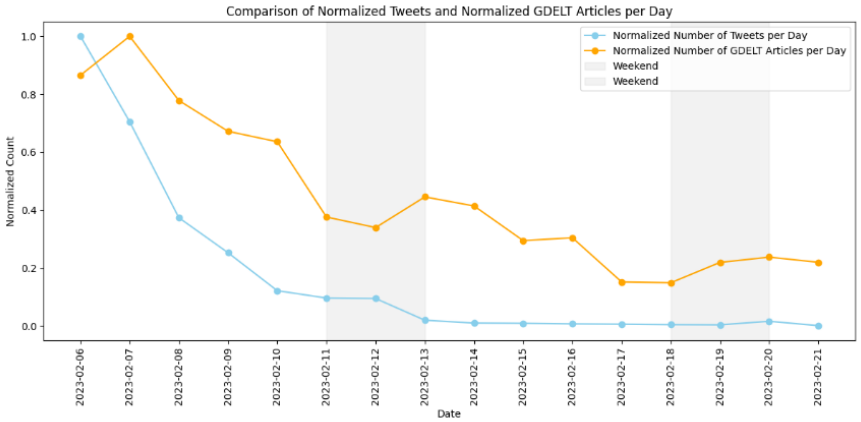

Figure 10:Comparison of social media and News Response of Turkey-Syria Earthquake

Appendix E: Trend Validation

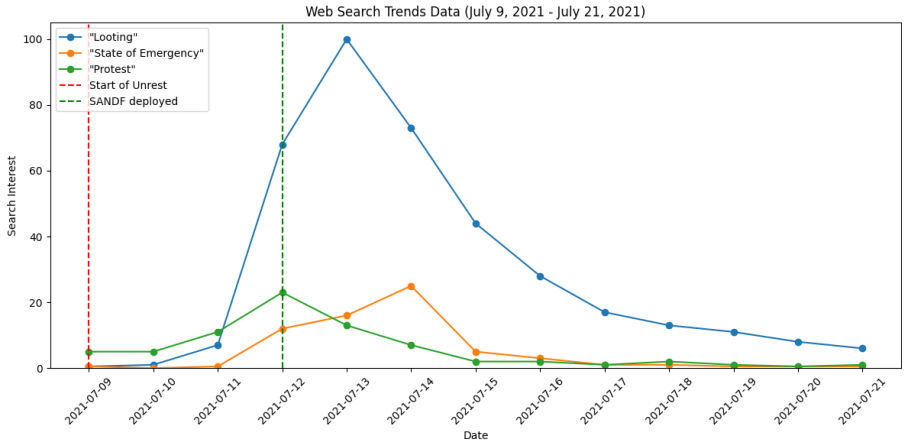

Figure 11: Trend Validation using Google Trends of 2021 South African Social Unrest

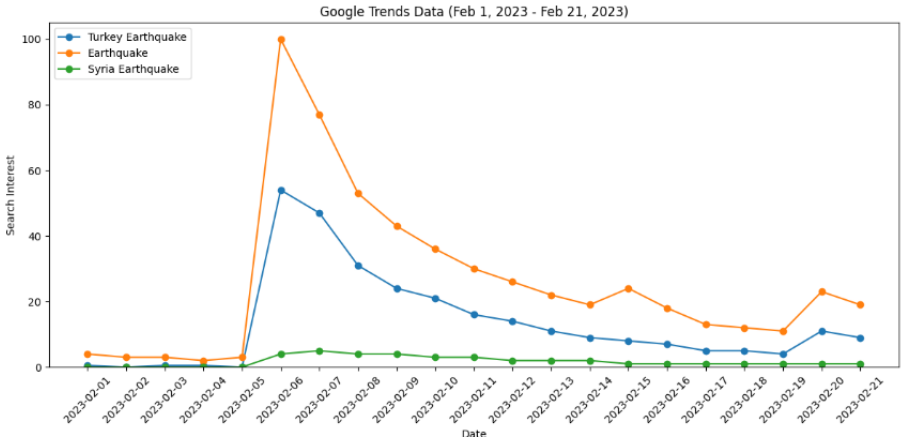

Figure 12: Trend Validation using Google Trends keywords of the Turkey-Syria Earthquake

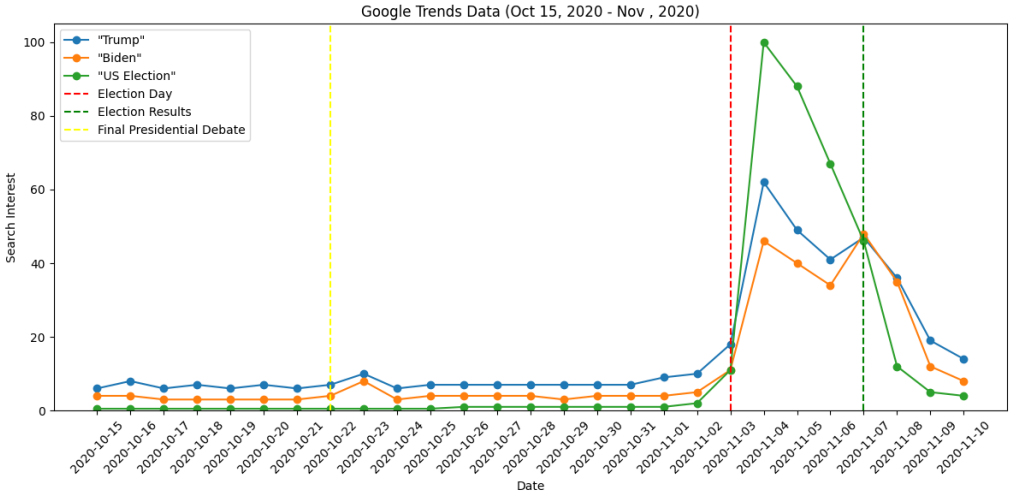

Figure 13: Trend Validation using Google Search of US 2020 Election

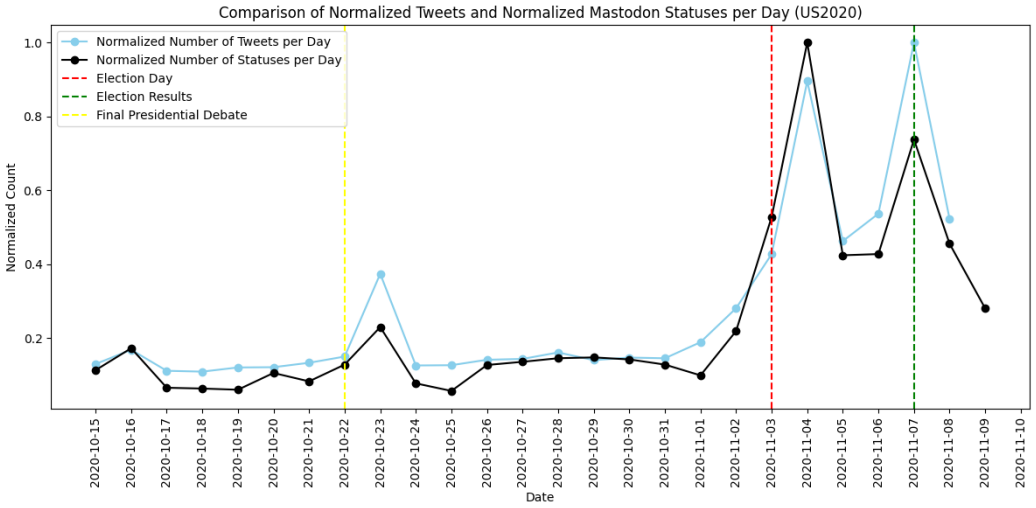

Figure 14: Trend Validation using Mastodon of the US Election in 2020

Appendix F: Event Archetypes

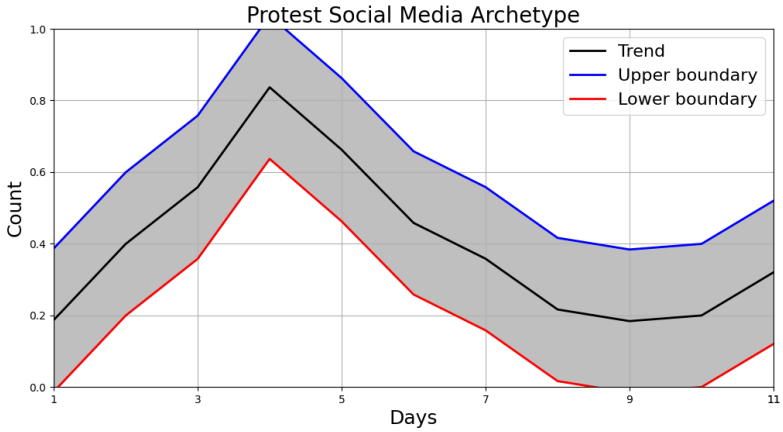

Figure 15: Event Archetype of Protest event type

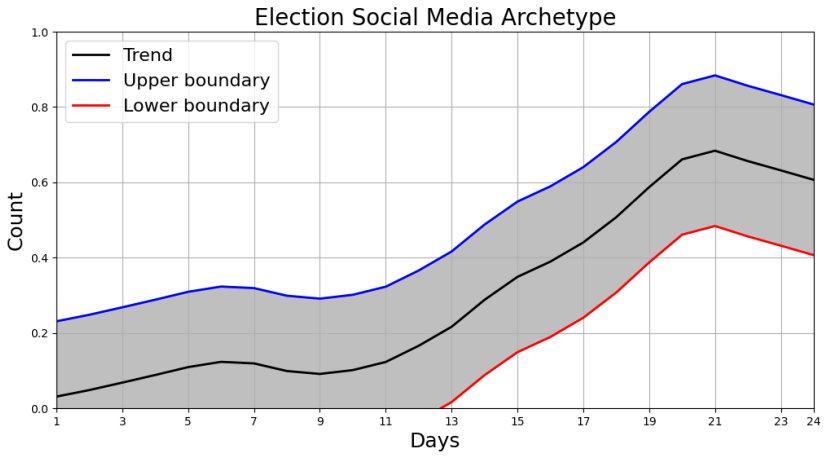

Figure 16: Event Archetype of Election event type

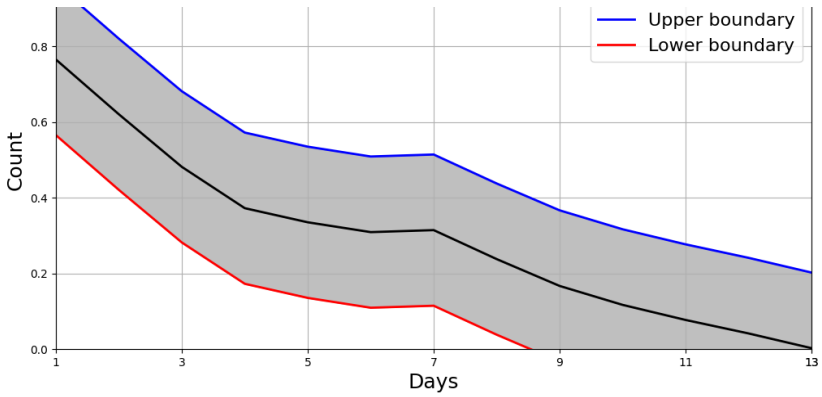

Figure 17: Event Archetype of Natural Disaster event type